Discrimination Avoidance Methods in Data Mining

Prajakta A. Soundankar

***Abstract*— Discrimination like privacy is a big issue when legal and ethical aspects of Data mining are considered. Most people don't like to be discriminated for their gender, religion, nationality, age and so on, especially when those attributes are needed for making decisions. Decisions like giving them a job, loan, insurance, etc. Hence it is highly desirable to discover such potential biases and eliminating them from the training data without harming their decision-making utility .Therefore antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining. Discrimination prevention consist inducing patterns which do not lead to discriminatory decisions even if the original training datasets are inherently biased. So By focusing on the discrimination prevention, we present a group of pre-processing discrimination prevention methods with different features of each approach and how**
These approaches deal with direct or indirect discrimination.

***Index Terms*— Antidiscrimination, data mining, direct and indirect discrimination prevention, rule protection, rule generalization,**
Privacy.

I. INTRODUCTION

Data mining is a rising and very essential technology for extracting useful knowledge hidden in large collections of data, especially human and social data sensed by the omnipresent technologies that support most human activities of our age. In fact, the new opportunities to extract knowledge and understand human and social complex phenomena increase conjointly with the risks of violation of fundamental human rights, such as privacy and non-discrimination. Privacy is the individual's right to choose freely what to do with one's own personal information, while discrimination is unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual merit. Human rights laws have deep concern about data protection [1] and they also prohibit discrimination [2, 3] against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy; and in a number of settings, like credit and insurance, personnel selection and wages, and access to public services. A wider social acceptance of a multitude of new Services and applications based on the knowledge discovery process can be achieved by preserving these great benefits of data mining within a privacy aware and discrimination aware technical ecosystem. Discrimination can be either direct discrimination which consists of rules or procedures that clearly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership or Indirect discrimination which could happen because of the availability of some background knowledge (rules).

Manuscript received January 07, 2015
Prajakta A. Soundankar Computer Department, MET BKC, Savitribai Phule Pune University, Nasik , India,

II. RELATED WORK

There is a more challenging issue apart from discrimination discovery, which is, preventing knowledge-based decision support systems from making discriminatory decisions (discrimination prevention) and when we want to prevent not only direct discrimination but also indirect discrimination or both at the same time it can become even more difficult. The way of eliminating discrimination and also to the phase of the data mining process in which discrimination prevention is done is related to the classification of discrimination prevention methods. Based on this criterion the discrimination prevention methods fall into three groups [4]:

- *Pre-processing*. Alter the source data in such a way that the discriminatory biases contained in the original data are removed and no unfair decision rule can be mined from the changed data and apply any of the standard data mining algorithms. The privacy preservation literature can be used to adapt the pre-processing approaches of data transformation and hierarchy based generalization. Along this line, [5], [6] perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased dataset.
- *In-processing*. Change the data mining algorithms in such a way that the resulting models will contain unfair decision rules [7], [8]. For example, an alternative approach to cleaning the discrimination from the original dataset is proposed in [7] where by the nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf relabeling approach. The in-processing discrimination prevention methods must rely on new special-purpose data mining algorithms; standard data mining algorithms cannot be used.

- *Post-processing*. Instead of cleaning the original dataset or changing the data mining algorithms work on or modify the resulting data mining algorithms.

III. Discrimination Avoidance Methods

The motive of all these methods is to transform the original data *DB* in a way as to remove direct or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data. The metrics that specify which records should be changed, how many records should be changed and how those records should be changed during data transformation are developed. The collection of data objects (records) and their attributes is a dataset. . Let *DB* be the original dataset. An item is an attribute along with its value, *e.g.* Race=black. An itemset, *i.e. X*, is a collection of one or more items, *e.g.* {Foreign worker=Yes, City=NYC}. A classification rule is an expression $x \rightarrow c$, where *C* is a class item (a yes/no decision), and *X* is an itemset containing no class item, *e.g.* {Foreign worker =Yes, City=NYC} Hire=no. *X* is called the premise of the rule. A frequent classification rule is a classification rule with a support or confidence greater than a specified lower bound. Let *DIs* be the set of predetermined discriminatory items in *DB* (*e.g.* DIs = {Foreign worker=Yes, Race=Black, Gender=Female}). The *support* of an itemset, $supp(x)$ is the fraction of records that contain the itemset *X*. We say that a rule $x \rightarrow c$ is *completely supported* by a record if both *X* and *C* appears in the record. The *confidence* of a classification rule, $conf(x \rightarrow c)$ measures how often the class item *C* appears in records that contain *X*. Hence, if $supp(x) > 0$

$$conf(x \rightarrow c) = \frac{supp(x, c)}{supp(x)}$$

A *frequent classification rule* is with a support or confidence greater than a specified lower bound. Let *FR* be the database of frequent Classification rules extracted from *DB*. The *negated itemset, i.e.* ~*X* is an itemset with the same attributes as *X*, butsuch that the attributes in ~*X* take any value except those taken by attributes in*X*. we use the ~ notation for itemsets with binary or categorical attributes. For a binary attribute, *e.g.* {Foreign worker=Yes/No}, if *X* is {Foreignworker=Yes}, then ~*X* is {Foreign worker=No}. Then, if *X* is binary, it can be converted to ~*X* and vice versa. However, for a categorical (non-binary) attribute, *e.g.* {Race= Black / White/Indian}, if *X* is {Race=Black}, then ~*X* is {Race=White} or {Race=Indian}. In this case, ~*X* can be converted to *X* without ambiguity, but the conversion of *X* into ~*X* is not uniquely defined, which we denote by ~*X*⟹ *X*. *Discriminatory attributes and itemset is* Attributes are classified as discriminatory according to the applicable anti-discrimination acts. Hence these attributes are regarded as discriminatory and the itemsets corresponding to them are called discriminatory itemsets. {Gender=Female, Race = Black} is just an example of a discriminatory itemset. Let DA_s be the set of predetermined discriminatory attributes in *DB* and DI_s be the set of predetermined discriminatory itemsets in *DB*. *Non-discriminatory attributes and itemsets* is the set of all the attributes in *DB* and I_s the set of all the itemsets in *DB*, then nDA_s (*i.e.* set of *nondiscriminatory attributes*) is $A_s - DA_s$ and nDI_s (*i.e.* set of *non-discriminatory itemsets*) is $I_s - DI_s$. An example of non-discriminatory itemset could be {Zip= 10451, City=NYC}.

A. Direct Discrimination Prevention Methods

The dataset of decision rules would be free of direct discrimination if it only contained PD rules that are *protective* or PD rules that are instances of at least one *nonredlining (legitimate)* PND rule. The proposed solution to prevent direct discrimination is based on this. Hence apply a suitable data transformation with minimum information loss in such a way that each *discriminatory* rule either becomes *protective* or an instance of a *non-redlining* PND rule. We call the first procedure direct rule protection and the second one rule generalization.

3.1.1 Direct Rule Protection (DRP)

To convert each *discriminatory* rule $r': A, B \rightarrow C$where A is a discriminatory itemset $(A \subseteq DI_s)$ *and* B *is non-discriminatory itemset* $(B \subseteq nDI_s)$, into a *protective* rule, two data transformation methods (DTM) could be applied. One method (DTM 1) changes the discriminatory itemset in some records (*e.g.* gender changed from male to female in the records with granted credits) and the other method (DTM 2) changes the class item in some records Table 13.1 shows the operation of these two methods.

Direct Rule Protection	
DTM1	$\sim A, B \rightarrow \sim C \Longrightarrow A, B \sim C$
DTM2	$\sim A, B \rightarrow \sim C \Longrightarrow \sim A, B, C$

Table 3.1 Data transformation methods for direct rule protection

$\sim C$ will be changed by modifying the value of the discriminatory itemset from $\sim A$ (Sex=Male) to A (Sex=Female) until *discriminatory* rule r': *A, B* $\rightarrow C$ *becomes protective* $(i.e\ elif: (r') < \alpha))$. Similar records are also chosen in DTM 2 with the difference that, instead of changing discriminatory itemsets, the class item is changed from ~*C* (grant credit) into *C* (deny credit) to make r' protective.

3.1.2 Rule Generalization

Another data transformation method for direct discrimination prevention is Rule Generalization which is based on each *discriminatory* rule $r': A, B \rightarrow C$ in the database of decision rules was an instance of at least one *nonredlining* PND rule $r: D, B \rightarrow C$ where *D* is a non-discriminatory itemset $(D \subseteq nDI_s)$, the dataset would be free of direct discrimination. To formalize this dependency among rules (i.e. *r'* is an instance of *r*), a PD classification rule *r'* is an instance of a PND rule $r$if rule r holds with the same or higher confidence, $conf(r: D, B \rightarrow C) conf(r': A, B \rightarrow C)$ and a case (record) satisfying discriminatory itemset *A* in context *B* satisfies legitimate itemset *D* as well, namely $conf(A, B \rightarrow D) = 1$. Based on this concept, a data transformation method (i.e. rule generalization) could be applied to transform each *discriminatory* rule $r': A, B \rightarrow C$ into an instance of a legitimate rule. Then, rule

generalization can achieved for discriminatory rules *r'* for which there is at least one *non-redlining* PND rule *r* by changing the class item in some records (e.g. from "Hire no" to "Hire yes" in the records of foreign and low-experienced people in NYC city). Table 3.2 shows the function of this method

.

Rule Generalization	
DTM	$A, B, \sim D \rightarrow C \Rightarrow A, B, \sim D \rightarrow \sim C$

Table3.2. Data transformation method for rule generalization

Table 13.2 shows that in DTM some records that support the rule *A, B, ~D → C* will change by modifying the value of class item from *C* (e.g. deny credit) into $\sim C$ (e.g. grant credit) until *discriminatory* rule $r': A, B \rightarrow c$ *becomes an instance of a non-redlining (legitimate)* PND rule $r: D, B \rightarrow C$.

3.1.3 Direct Rule Protection and Rule Generalization

Though all *discriminatory* rules may not be transformed by Rule generalization and, also rule generalization cannot be used alone for direct discrimination prevention and must be combined with direct rule protection. When applying both rule generalization and direct rule protection, *discriminatory* rules are divided into two groups:

- *Discriminatory* rules r' for which there is at least one non-redlining PND rule r such that r' could be an instance of r . For these rules, rule generalization is performed unless direct rule protection requires less data transformation (in which case direct rule protection is used).
- *Discriminatory* rules *r'* such that there is no such PND rule. For these rules, directrule protection (DTM 1 or DTM 2) is used.

3.2.2 Indirect Discrimination Prevention Methods

To avoid indirect discrimination is based on the fact that the dataset of decision rules would be free of indirect discrimination if it contained no *redlining* rules. It is achieved by a suitable data transformation with minimum information loss should be applied in such a way that *redlining* rules are converted to *non-redlining* rules. This procedure is called indirect rule protection (IRP). In order to turn a *redlining* rule $r: D, B \rightarrow C$ where *D* is a non-discriminatory itemset that is highly correlated to the discriminatory itemset *A*, into a *nonredlining* rule based on the indirect discriminatory measure (elb) two data transformation methods could be applied, like the ones for direct rule protection. One method (DTM 1) changes the discriminatory itemset in some records (*e.g.* from non-foreign worker to foreign worker in the records of hired people in NYC city with Zip≠ 10451) and the other method (DTM 2) changes the class item in some records (e.g. from "Hire yes" to "Hire no" in the records of non-foreign worker of people in NYC city with Zip≠ 10451). Table 3.3 shows the operation of these two methods.

Indirect Rule Protection	
DTM1	$\sim A, B, \sim D \rightarrow \sim C \Rightarrow A, B, \sim D \rightarrow \sim C$
DTM2	$\sim A, B, \sim D \rightarrow \sim C \Rightarrow \sim A, B, \sim D \rightarrow C$

TABLE 3.3 DATA TRANSFORMATION METHODS FOR INDIRECT RULE PROTECTION

Table 3.3 shows that in DTM 1 some records in the original data that support the rule *~A, B, ~D → ~C* will be changed by modifying the value of the discriminatory itemset from *~A* (Sex=Male) into *A* (Sex=Female) in these records until the *redlining* rule $r: D, B \rightarrow C$ becomes *non-redlining* $(i.e\ elb(r) < \alpha)$ .Similar records are also chosen in DTM 2 with the difference that, instead of changing discriminatory itemsets, the class item is changed from *~C* (e.g. grant credit) into *C* (e.g. deny credit) in these records to make *r non-redlining*. The difference between the DRP and IRP methods shown in Tables 1 and 3 is about the set of records chosen for transformation. As shown in Table 3, in IRP the chosen records should not satisfy the *D* itemset (chosen records are those with $\sim A, B, \sim D \rightarrow \sim C$) whereas DRP does not care about *D* at all (chosen records are those with $\sim A, B \rightarrow \sim C$).

IV. CONCLUSION

When legal and ethical aspects of data mining are considered Discrimination is a Big and Important issue since most people do not want to be discriminated on account of their gender, religion, nationality, age, and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. The motive of this paper was to develop a new preprocessing discrimination prevention methodology which includes different data transformation methods which can prevent direct discrimination, indirect discrimination or both of them at the same time. In order to attain this objective, initial step is to measure discrimination identifying categories and groups of individuals who have been directly and/or indirectly discriminated in the decision-making processes; the second step is to transform data in the proper way to remove all those discriminatory biases. Finally, without seriously damaging data quality discrimination free data models can be produced from the transformed data. The experimental results described, demonstrate that the proposed techniques are quite successful in both goals of removing discrimination and preserving data quality.

ACKNOWLEDGMENT

The author wish to thank MET's Institute of Engineering Bhujbal Knowledge City Nasik, HOD of computer department, guide and parents for supporting and

motivating for this work because without their blessing this was not possible.

REFERENCES

[1] European Union Legislation. Directive 95/46/EC, 1995.

[2] Australian Legislation. (a) Equal Opportunity Act Victoria State, (b) Anti-Discrimination Act -Queensland State, 2008. http://www.austlii.edu.au/.

[3] European Union Legislation, (a) Race Equality Directive,2000/43/EC, 2000; (b) Employment Equality Directive, 2000/78/EC, 2000; (c) Equal Treatment of Persons, European Parliament legislative resolution, P6 TA(2009)0211, 2009.

[4] S. Ruggieri, D. Pedreschi and F. Turini, "Data mining for discrimination discovery", ACM Transactions on Knowledge Discovery from Data, 4(2) Article9, ACM, 2010.

[5] F. Kamiran and T. Calders, "Classification without discrimination", Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009). IEEE, 2009.

[6] F. Kamiran and T. Calders, "Classification with no discrimination by preferentialsampling", Proc. of the 19th Machine Learning conference of Belgium and The Netherlands, 2010.

[7] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification", Data Mining and Knowledge Discovery, 21(2):277-292,2010.

[8] F. Kamiran, T. Calders and M. Pechenizkiy, "Discrimination aware decision tree learning", Proc. of the IEEE International Conference on Data Mining (ICDM2010), pp. 869-874. ICDM, 2010.

[9] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09), pp. 581-592, 2009.

[10] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating Induction and Deduction for Finding Evidence of Discrimination," Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09), pp. 157-166, 2009.

[11] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.

[12] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases," Proc. ACM Int'l Conf. Management of Data(SIGMOD '10), pp. 1127-1130, 2010.